# NeRF as Pretraining at Scale: Generalizable 3D-Aware Semantic Representation Learning from View Prediction

Wenyan Cong[1], Hanxue Liang[2], Zhiwen Fan[1], Peihao Wang[1], Yifan Jiang[1],
Dejia Xu[1], A. Cengiz Oztireli[2], Zhangyang Wang[1]
[1]University of Texas at Austin [2]University of Cambridge
{wycong, zhiwenfan, peihaowang, yifanjiang97, dejia, atlaswang}@utexas.edu, {hl589, aco41}@cam.ac.uk

## Abstract

*Cross-scene generalizable NeRF models, which could directly synthesize novel views using several source views of unseen scenes, are gaining prominence in the NeRF field. Discovering the potential signal of emerging capabilities in existing methods, we draw a parallel between BERT's "drop-and-predict" Masked Language Model (MLM) pretraining and novel view synthesis (NVS) in generalizable NeRF. In this work, we pioneer the scaling up of NVS as an effective pretraining strategy in a multi-view context. To bolster generalizability in pretraining, we incorporate a large-scale, minimally annotated dataset and proportionally increase the model size, revealing a neural scaling law akin to that observed in BERT. We also introduce innovative hardness-aware training techniques to enhance robust feature learning. Our model, named "NPS", demonstrates remarkable generalizability in both zero-shot and few-shot novel view synthesis. It further shows emergent capabilities in downstream tasks like few-shot multi-view semantic segmentation and depth estimation. Significantly, NPS reduces the necessity of training separate models for each task, underlining its versatility and efficiency. This approach sets a new precedent in the NeRF field, and highlights the vast possibilities opened up by scaling up generalizable novel view synthesis.*

## 1. Introduction

Neural Radiance Field [23] (NeRF) has achieved remarkable success on synthesizing novel views given multi-view source images. Recently, generalizable NeRF models [4, 31, 38, 39, 45, 49] mark a shift from per-scene fitting to efficient transformer-based 'feedforward' generation. Among them, [4, 38] demonstrated two interesting properties: *improved few-shot learning ability*, indicating enhanced generalization capabilities in data-limited scenarios, and *the emergence of depth awareness* through self-attention without explicit depth training.

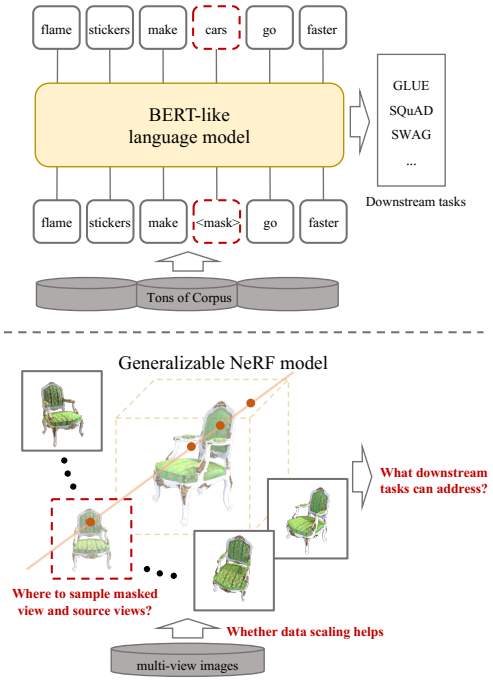



Figure 1. If we treat each view as a "word token", the training of the generalizable NeRF model is similar to Masked Language Model (MLM) pretraining of BERT. However using such MLM training in multi-view scenarios is challenging in deciding 1) whether data scaling helps, 2) where to sample masked view and source views, and 3) what downstream tasks can address.

Exploring these properties as emergent capabilities, which is typical of Large Language Models (LLMs) developed through extensive training, reveals significant parallels between the training methodologies of generalizable NeRF and LLMs like BERT [6]. As in Figure 1, in novel view synthesis, if we treat each view as a "word token", predicting omitted target view using source views aligns

with the "drop-and-predict" mechanism of the "Masked Language Model" (MLM) adopted in BERT's pretraining. Such a simple meta-training strategy in language tokens has successfully endowed BERT with robust few-shot learning and emergent behavior on multiple language downstream tasks. Therefore, the parallel prompts a compelling question: Could the training approach of novel view synthesis used in generalizable NeRF model, serve as a mirror of MLM concept in vision domain? It means that after sufficient large-scale training, the model can easily adapt to multiple downstream vision tasks, in a manner akin to BERT's widespread use in natural language understanding.

The concept of MLM is not new to vision domain. A similar "drop-and-predict" pretraining strategy has been developed to reconstruct the masked image patches in Masked Autoencoder (MAE) [13] in 2D vision, or to recover masked point cloud based on partial observations in 3D vision [17, 24]. However, these extensions primarily focus on extracting image-level features or 3D small-scale local neighboring information. In contrast, novel view synthesis, a distinct instantiation of MLM on a novel multi-view level, leverages a broader range of views, rather than just a small, localized area, to predict new views. Despite its potential to significantly enhance understanding of geometric structures and global scene context, as has also been noted in [4, 38], this multi-view level MLM model remains largely under-explored.

Despite the promise, challenges persist to develop novel view synthesis as an effective multi-view level MLM pretraining: ❶ *Whether data scaling helps*: The emergent few-shot learning ability of BERT relies heavily on large-scale training, a scale not yet fully explored in current generalizable NeRF models. The curse of dimensionality is a major issue with multi-view data. It remains unclear whether simply scaling up the volume of data will enable the emergence of advanced capabilities, as seen in 1D LLMs and 2D vision models. Meanwhile, the amount of data required for effective generalizable pretraining also remains unknown. ❷ *Where to sample masked view and source views*: In language models, each word token has a well-defined position within a sequence. Similarly in 2D vision, an analogical structure exists in the form of a 2D grid. However, in multi-view scenarios, defining the position of the masked "view token" is more complex due to the additional dimensions involved. Meanwhile, deciding which source views to use for predicting adds complexity to the task. ❸ *What downstream tasks can address*: In the context of multi-view pretraining, apart from depth estimation, which has shown potential in preliminary studies [28], it's uncertain what other downstream tasks could be effectively addressed. For instance, the feasibility of performing semantic segmentation without prior exposure to semantic information is yet to be determined.

In this paper, we tackle the outlined challenges through three main explorations. First, **Model and Data Scaling**. We employ the state-of-the-art generalizable NeRF model, GNT [38], as the foundation framework for investigating scalability. The training data is augmented through the integration of the latest OmniObject3D dataset [41]. We discover that continuous data scaling necessitates a proportional increase in model size, revealing a neural scaling law in generalizable NeRF similar to that observed in BERT.

Second, **Hardness-Aware Sampling Strategies**. To enhance the model's generalizability, we adopt hardness-aware sampling strategies during pretraining. This involves progressively decreasing the number of source views, increasing the distance between target and source views, and using non-regular sampling patterns. Such strategies gradually add complexity to the training process, fostering the model's capacity to learn robust features across diverse data scenarios.

Third, **Exploring NeRF-Pretrained Features**. Besides directly evaluating the performance of the generalizable NeRF pretraining through novel view synthesis, our primary interest lies in uncovering the emergent capabilities of these pretrained features. This is achieved by applying them to various downstream tasks. A notable discovery is that our model excels in few-shot multi-view semantic segmentation, even without specific pretraining for semantic information. This surprising performance surpasses that of models explicitly designed for this task, suggesting that NeRF pretrained models possess an innate ability to comprehend and bridge the semantic gaps through the analysis of multi-view 3D world observations.

Our main contributions could be summarized as follows:

- We introduce NeRF as Pretraining at Scale (NPS) and for the first time, demonstrate that paralleling the MLM pretraining in BERT, generalizable novel view synthesis can be effectively scaled up as a pretraining task in multi-view scenario, pioneering the application of multi-view MLM techniques and setting a precedent for future research.
- To scale up generalizable NeRF, we incorporate a larger dataset with minimal annotations and scale up the model size accordingly in the pretraining stage, revealing a neural scaling law in generalizable NeRF akin to that observed in BERT. We also introduce hardness-aware training techniques to guarantee robust feature learning.
- Going beyond assessing generalizable NeRF pretraining through both zero-shot and few-shot novel view synthesis, we crucially evaluate the emergent capabilities of NeRF pretrained features in downstream tasks, including multi-view semantic segmentation and depth estimation in a few-shot setting. NPS could outperform specialized models, underscoring the potential of NeRF pretrained models to bridge the semantic gap through multi-view observations alone.

## 2. Related Works

**Generalizable NeRF** NeRF needs to be trained from scratch for each target scene, with no prior knowledge extracted and shared among scenes, which severely reduces the efficiency of scene reconstruction and novel-view synthesis. We can differentiate between two key approaches for generalizing across scenes. One line of work [14, 34, 48] is often implemented via local conditioning, where the coordinate input to the scene representation MLP is concatenated with a locally varying feature vector, stored in a discrete scene representation, such as a voxel grid [26]. PixelNeRF [48] leverages the volume rendering framework, where encoded image features are aggregated over multiple views, and an MLP produces color and density fields that are rendered as in NeRF. GRF [34] uses a similar framework, with an additional attention module that reasons about the visibility of 3D points in the different sampled input images. Recent advances [4, 30, 32, 38, 39] adopt transformer-based networks with epipolar constraints for novel view synthesis in "feedforward" fashion. IBRNet [39] introduces transformer networks across the ray samples that reason about visibility. Light Field Networks [30] optimize an MLP to directly encode the mapping from an input ray to an output color (the scene's light field), enabling single-evaluation rendering. GNT and its variant [4, 38] use a pure, unified transformer-based architecture to reconstruct NeRFs, where the view transformer aggregates information from epipolar lines on the neighboring views, and ray transformer aggregates the feature along the ray to obtain the final rendering.

**MLM Pretraining in Language and Vision** Masked Language Model (MLM) pretraining has emerged as a pivotal technique, synergizing the domains of natural language processing [6, 20, 29, 35, 43] and computer vision [9, 13, 17, 24, 33, 40]. This approach, exemplified by models like BERT [6] in language and its visual counterparts, leverages the concept of masking key elements in data (text or visual features) and learning to predict them. In computer vision, this technique has been adapted to enhance understanding in both 2D and 3D spaces. In 2D vision [9, 13], MLM pretraining aids in extracting intricate features from images, facilitating tasks like object detection and image segmentation. In the 3D domain [17, 24], this approach is instrumental in understanding spatial relationships and depth perception, crucial for applications in robotics and augmented reality. Compared to those models, our approach is the first model to develop novel view synthesis as a new instantiation of MLM at the multi-view level, and explores its emergent capacities for geometric and semantic awareness after large-scale training.

**Multi-view Feature Learning with NeRF** The success of the NeRF in novel view synthesis has inspired researchers to develop it on multi-view feature learning and image understanding tasks [8, 12, 18, 28, 37, 47, 50]. VolRecon [28] proposes a new pipeline for generalizable implicit reconstruction that produce detailed surfaces. NeRF-SOS [8] conducts exploration of self-supervised learning for object segmentation using NeRF for complex real-world scenes. [12, 18, 37, 50] learn high-level semantic understanding with the NeRF structure. [47] proposes a unified framework for learning generalizable NeRFs from distilling pretrained 2D vision foundation models. Different from existing methods, our NPS explores the potential of scaling simple generalizable novel view synthesis as a pretraining task and demonstrates its representation learning capability on downstream tasks.

**Large-Scale Datasets** The acquisition of large-scale, realistic 3D datasets is costly and challenging. Commonly used datasets like ShapeNet [1] and ModelNet40 [42] primarily consist of synthetic CAD models. Recent contributions such as 3D-FUTURE [11] and ABO [3] introduce high-quality CAD models, yet the difference between synthetic and real-world objects remains a challenge. Photo-realistic datasets like DTU [15] and BlendedMVS [46] are limited in scale and category diversity. ScanObjectNN [36] offers real-world point clouds but with incomplete and noisy data. Google Scanned Objects [7] and AKB-48 [19], while detailed, have a narrow semantic range. CO3D [27] offers multi-view images of 15k objects, yet it suffers from varied quality and limited object categories. The latest OmniObject3D dataset excels with 6k professionally scanned 3D objects accompanied by high-quality multi-view photographs across 190 categories, making it versatile for NeRF models.

## 3. Method

**Overview.** As shown in Figure 2, our whole framework is composed of two stages: pretraining on multi-view images and fine-tuning on downstream tasks. During pretraining stage, the model is trained with novel view synthesis task. During fine-tuning stage, we add task-specific head to the pretrained NPS model, and fine-tune the whole model with both novel view synthesis and downstream task. A distinctive advantage of NPS model lies in its unified architecture, which is seamlessly adaptable across diverse tasks. It requires minimal customization between the pretrained and final downstream architectures.

### 3.1. Preliminary

The Generalizable NeRF Transformer (GNT) [38] is a state-of-the-art, transformer-based architecture for generalizable NeRF. It comprises two stages: the "view transformer" and the "ray transformer". The view transformer first aggregates

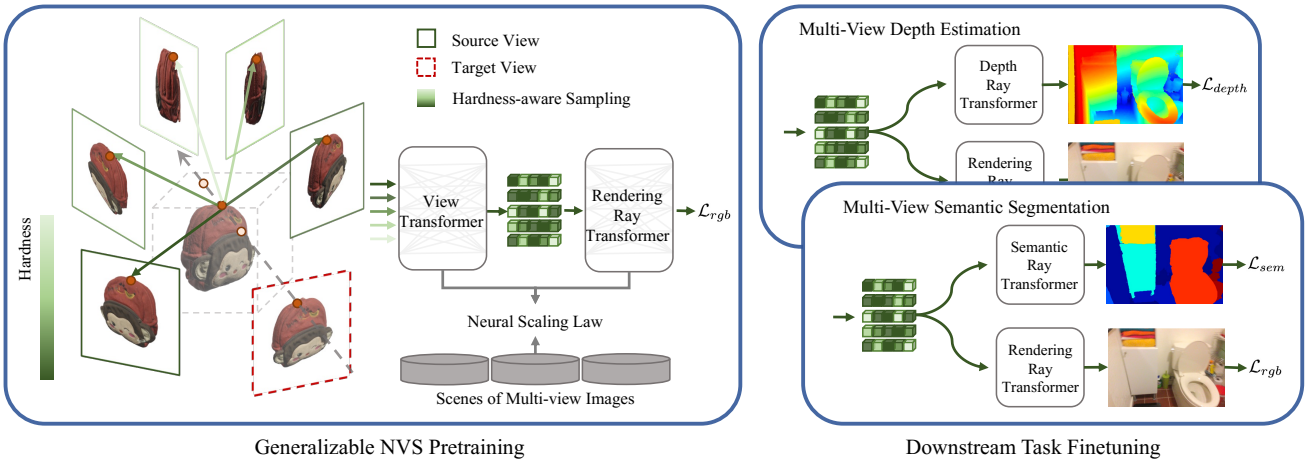



Figure 2. Overall pretraining and fine-tuning procedures for NPS. Apart from output layers, the same architectures are used in both pretraining and fine-tuning. The same pretrained model parameters are used to initialize models for different downstream tasks. During fine-tuning, all parameters are fine-tuned.

information across epipolar lines from neighboring views to predict aligned features for each 3D point. Given $N$ source images $\{\mathbf{I}_i \in \mathbb{R}^{H\times W\times 3}\}_{i=1}^{N}$, for each point $\boldsymbol{x} \in \mathbb{R}^3$ on a ray emitted from the target view, it operates as follows:

$$\mathcal{F}(\boldsymbol{x}, \boldsymbol{\theta}) = \text{V-Trans}(\mathbf{F}(\Pi_1(\boldsymbol{x})), ..., \mathbf{F}(\Pi_N(\boldsymbol{x}))), \quad (1)$$

where $\Pi_i(\boldsymbol{x})$ projects point $\boldsymbol{x}$ onto the $i$-th image plane $\mathbf{I}_i$, and $\mathbf{F}$, a small U-Net-based CNN, interpolates features at the projected image point. These multi-view aggregated features are then processed by the ray transformer. Then an MLP is employed to map them to RGB colors:

$$\mathcal{C}(r) = \text{MLP} \circ \text{R-Trans}(\mathcal{F}(\boldsymbol{x}_1, \boldsymbol{\theta}), ..., \mathcal{F}(\boldsymbol{x}_M, \boldsymbol{\theta})). \quad (2)$$

Here, $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_M\}$ are points along ray $r$. We choose GNT as our backbone due to its impressive performance and its inherently scalable nature, though our pretraining techniques can be generalized to other transformer-based NeRFs [4, 16, 25, 32].

### 3.2. Pretraining NPS

**Scaling Pretraining Data** The emergent learning capabilities of BERT rely extensively on large-scale training, a magnitude not fully explored in current generalizable NeRF models. Previous approaches have predominantly operated within the confines of small-scale experimental setups. Standard datasets such as the LLFF [22], Google Scanned Objects [7], Spaces dataset [10], RealEstate10K [51], and DTU [15] have been the cornerstones of training data for these models.

The recent emergence of large 3D datasets has significantly expanded the scale and variety of available training data. Among them, the Omniobject3D dataset is a prime example, featuring 6,000 professionally scanned 3D objects accompanied by high-quality multi-view images across 190 categories. In an effort to mirror the large-scale training approach of BERT and investigate potential emergent capabilities, we integrate the Omniobject3D dataset into our pretraining regime. This represents a notable departure from the traditionally constrained scope of generalizable NeRF models.

**Scaling Model Size** NPS builds upon the current SOTA of generalizable NeRF, GNT [38], a pure transformer-based network able to synthesize novel views from a set of source views. The natural scalability of GNT's transformer network facilitates experiments with varying model sizes, and as we expand the training data, we observe a parallel need to scale up the model size.

Through systematic experiments, we find that the performance of GNT improves significantly with the concurrent scaling of data and model size, suggesting that there may be a deeper, underlying principle at play. This trend, depicted in Figure 3, resembles neural scaling laws observed in language models like BERT, which not only reinforce the importance of large-scale data in training but also highlight the necessity of larger model capacities to fully exploit this data. The existence of a neural scaling law for generalizable NeRF pretraining opens up exciting avenues for future research. It encourages us to consider how we might further optimize the balance between data scale and model size to push the boundaries of what is achievable in 3D vision and rendering.

**Hardness-aware Training Strategies** We observe that large-scale multi-view datasets exhibit a clustering nature due to different views of the same scene, and even differ-

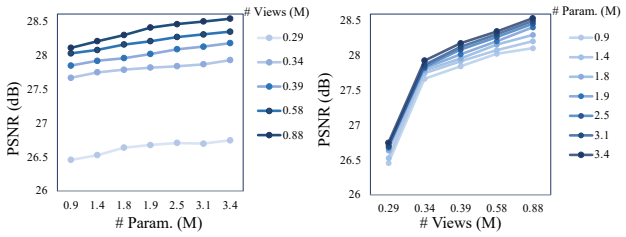



Figure 3. Neural scaling law founded in the pretraining of generalizable NeRF with scaling pretraining data and scaling model size.

ent scenes, will bear natural scene similarity. These characteristics can significantly complicate the training process, as the models must learn to generalize across a wide range of data distributions. To tackle this issue, we propose a set of hardness-aware training techniques that incrementally introduce complexity during the training process, to enhance the model's ability to learn robust features across diverse data scenarios. As shown in Figure 4, this is achieved through the following strategies:

1) Gradually Decreasing Source Views: In the initial stages of training, the model is provided with a larger number of source views (typically 10, as in the standard GNT setup). Over time, we reduce this number to as few as 2 source views, forcing the model to make more accurate predictions with less information, which is a harder task that promotes better generalization, thus helping the network generalize to sparse scenes and few-shot settings.

2) Gradually Increasing View Distance:Given the camera pose of the target view and source view, the view distance is computed based on the camera locations. We start by training the model with source views that are close to the target view, which provides a simpler learning task. As training progresses, we increase the distance between the source and target views, thereby introducing more complexity and encouraging the model to learn more generalizable features.

3) Non-Regular Sampling of Source Views: Instead of selecting the nearest source views, which may lead to overfitting on specific view arrangements, we employ a uniform sampling strategy. This method ensures that the model is exposed to a more varied set of perspectives, further enhancing its ability to generalize.

The implementation of these hardness-aware training techniques is carefully calibrated to maintain a balance between the model's capacity and the complexity of the data. By incrementally adjusting the training difficulty, we ensure that the model is not overwhelmed by the complexity at the outset, which could hinder learning.

By employing these strategies, we aim to harden the training process in a controlled manner. The expected outcome is a model that not only performs well on the training data but also exhibits improved generalization to new, unseen data.

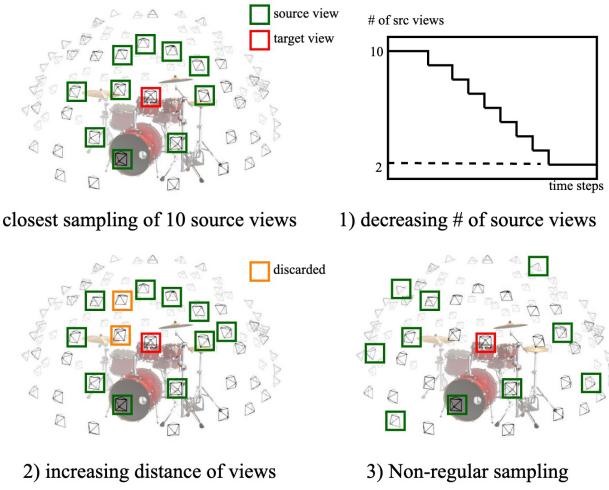



Figure 4. Different from the sampling strategy of GNT (top left corner), our hardness-aware sampling strategies could harden the training process and force the model to adapt and learn more effectively.

### 3.3. Fine-tuning NPS

The ray transformer in GNT plays a crucial role in creating a flexible framework suitable for downstream tasks in multi-view settings. Its ability to aggregate features along a ray, utilizing the self-attention mechanism, provides NPS with the versatility to handle outputs from diverse domains. This functionality enables the use of distinct ray transformers for different downstream tasks, eliminating the necessity for extensive task-specific architectural changes and making it possible to render novel views along with their corresponding labels across various domains, such as depth maps and semantic maps. By leveraging multi-view source images, NPS circumvents the necessity for ground truth input from the target view. This approach not only simplifies the training process but also enhances the robustness of the model to the intricacies inherent in 3D vision tasks.

For each unique downstream task, NPS's architecture is designed to be augmented with a task-specific ray transformer block, which is more lightweight compared to the rendering ray transformer. This design choice ensures that, despite its adaptability, the computational load does not significantly increase, maintaining efficiency in the framework's operation. The block could be seamlessly integrated into the existing framework, allowing for a unified approach to fine-tuning. The process involves end-to-end training, where all parameters of NPS are adjusted simultaneously to optimize performance for novel view synthesis and the downstream task at hand.

# 4. Experiments

## 4.1. Pretraining Cross-Scene Generalizaiton

### 4.1.1 Implementation Details

**Training Details** As in the right subfigure in Figure 3, we observe a convergence of performance after scaling up the model size. To balance between the model size and computational resources and facilitate the training process, we opted for a scaled-up version of the GNT model with 1.9M parameters. This model configuration includes 8 blocks for both view transformer and ray transformer, and a latent dimension of 768 for the MLP layer. Please refer to our supplementary material for more details.

**Setting** Following [4, 38], we compare our NPS with state-of-the-art generalizable NeRF under both zero-shot and few-shot novel view synthesis. For zero-shot setting, the pretrained model is directly evaluated on an unseen scene for novel view synthesis. For few-shot setting, the pretrained model is first fine-tuned with a few observed views (as few as 3) from the target unseen scene, and then applied to the target scene.

**Baselines** To evaluate the novel view synthesis performance, we compare our method with existing generalizable NeRF models, including PixelNeRF [49], MVSNeRF [2], IBRNet [39], GNT [38], GPNR [31], and GNT-MOVE [4].

**Datasets** (1) During the pretraining stage, besides the five training datasets of GNT, including Google Scanned Object [7], RealEstate10K [51], Spaces dataset [10], and real scenes from handheld cellphone captures [22, 39], used in [4, 38, 39], we also incorporate the multi-view images from large-scale dataset Omniobject3D [41], which comprises 6,000 3D objects scanned by professional devices in 190 categories. (2) Testing Datasets are the common NeRF benchmarks including Local Light Field Fusion (LLFF) [22] and NeRF Synthetic dataset [23].

### 4.1.2 Zero-Shot Generalization

Note that we incorporate the hardness-aware training strategies in the training, which pose more training challenges for generative novel view synthesis. For example, in the later training stage, the model is enforced to reconstruct a novel view from only two far-away source views, which may initially impede performance. However, as evidenced in Table 1, our NPS model still excels in zero-shot generalization, particularly on the NeRF Synthetic dataset. This performance underscores the benefit of scaling up pretraining for enhanced generalization. This trend is also evident in the qualitative results displayed in Figure 5.

| Models | Local Light Field Fusion (LLFF) | | | NeRF Synthetic | | |
|---|---|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ |
| PixelNeRF | 18.66 | 0.588 | 0.463 | 22.65 | 0.808 | 0.202 |
| MVSNeRF | 21.18 | 0.691 | 0.301 | 25.15 | 0.853 | 0.159 |
| IBRNet | 25.17 | 0.813 | 0.200 | 26.73 | 0.908 | 0.101 |
| GPNR | 25.72 | 0.880 | 0.175 | 26.48 | 0.944 | 0.091 |
| GNT | 25.86 | 0.867 | 0.116 | 27.29 | 0.937 | 0.056 |
| GNT-MOVE | 26.02 | 0.869 | 0.043 | 27.47 | 0.940 | 0.056 |
| Ours | **26.17** | **0.873** | **0.104** | **28.41** | **0.951** | **0.052** |

Table 1. Zero-shot generalization performance on NeRF Synthetic dataset and LLFF dataset.

### 4.1.3 Few-Shot Generalization

We fine-tune our pretrained models on two datasets: the LLFF dataset, using subsets of 3, 6, and 10 forward-facing images, and the NeRF Synthetic dataset, using 6 or 12 360° images. During the inference phase, we utilize the same views from the finetuning stage as source views, which not only ensures fairness in our evaluation but also presents a significant challenge, testing the model's capability to generalize from limited data. The results are reported in Table 2. Our NPS outperforms all the baselines in all the few-shot settings. It is noteworthy that NPS is especially helpful in sparse scenes: 3-shot, 6-shot on LLFF and 6-shot on Blender, which benefits from the hardness training strategies in the pretraining stage.

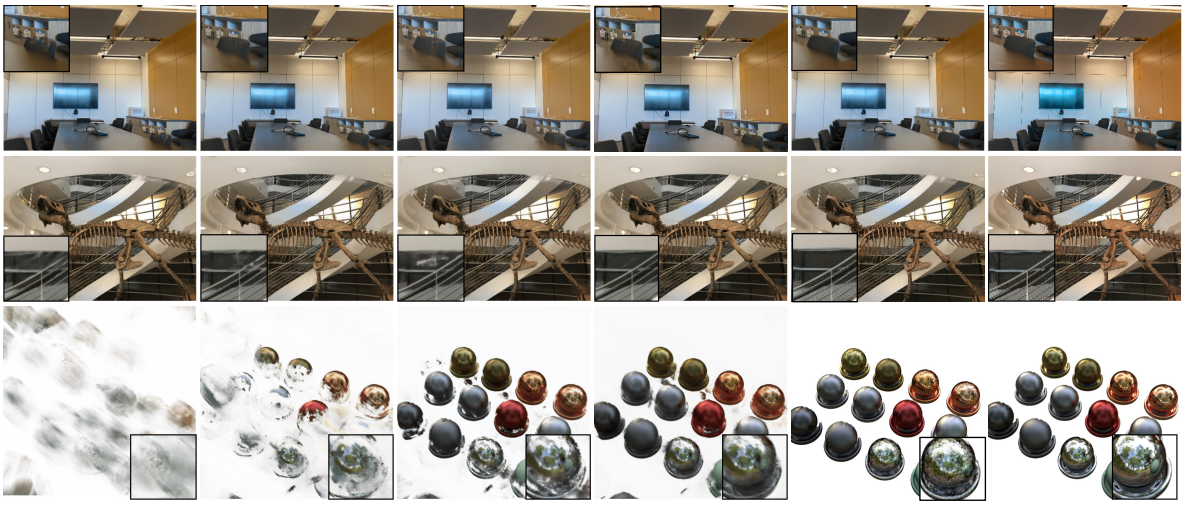
## 4.2. Few-Shot Downstream Task Generalization

Our experimental approach evaluates pretrained NeRF features in downstream tasks, particularly under few-shot conditions, using labels from only a few scenes. This choice is motivated by the practical difficulties in acquiring extensive annotations. Such a setup more accurately reflects real-world conditions and facilitates a thorough assessment of the pretrained NeRF models' emergent capabilities in data-scarce environments.

**Datasets** To evaluate downstream tasks, we use a large labeled RGB-D dataset, ScanNet [5], which contains 2.5M views in 1513 scenes annotated with 3D camera poses, surface reconstructions, and semantic segmentation. We choose 60 different scenes as training datasets and 10 unseen novel scenes as test datasets to evaluate generalizability in real data.

### 4.2.1 Multi-View Semantic Segmentation

**Baselines** The most related baseline is Semantic-Ray [18], a generalizable semantic field in real-world scenes. To construct more baselines to compare the downstream performance, following [18], we add the semantic head same as the one in Semantic-NeRF on generalizable NeRF models, including MVSNeRF [2] and NeuRay [21]. In addition, we also compare with



(a) IBRNet (b) GPNR (c) GNT (d) GNT-MOVE (e) Ours (f) GT

Figure 5. Qualitative results for the unseen cross-scene rendering.

| Models | Local Light Field Fusion (LLFF) | | | | | | | | | NeRF Synthetic | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 3-shot | | | 6-shot | | | 10-shot | | | 6-shot | | | 12-shot | | |
| | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ |
| PixelNeRF | 17.54 | 0.543 | 0.502 | 19.00 | 0.721 | 0.496 | 20.01 | 0.755 | 0.333 | 19.13 | 0.783 | 0.250 | 21.90 | 0.849 | 0.173 |
| MVSNeRF | 17.05 | 0.486 | 0.480 | 20.50 | 0.594 | 0.384 | 22.54 | 0.673 | 0.309 | 16.74 | 0.781 | 0.263 | 22.06 | 0.844 | 0.185 |
| IBRNet | 16.89 | 0.539 | 0.458 | 20.61 | 0.686 | 0.316 | 23.52 | 0.789 | 0.226 | 18.17 | 0.812 | 0.234 | 24.69 | 0.895 | 0.120 |
| GNT | 19.58 | 0.653 | 0.279 | 22.36 | 0.766 | 0.189 | 24.14 | 0.834 | 0.133 | 22.39 | 0.856 | 0.139 | 25.25 | 0.901 | 0.088 |
| GNT-MOVE | 19.71 | 0.666 | 0.270 | 22.53 | 0.774 | 0.184 | 24.61 | 0.837 | 0.132 | 22.53 | 0.871 | 0.116 | 25.85 | 0.915 | 0.074 |
| Ours | **20.02** | **0.673** | **0.263** | **22.78** | **0.789** | **0.174** | **24.79** | **0.866** | **0.111** | **22.78** | **0.880** | **0.111** | **25.97** | **0.926** | **0.068** |

Table 2. Comparison of NPS with existing generalizable NeRF methods in a few-shot setting.

Semantic-NeRF [50], which is not generalizable and needs per-scene optimization.

**Metrics** We use mean Intersection-over-Union (mIoU), average accuracy, and total accuracy for evaluating segmentation quality, and PSNR, SSIM, and LPIPS metrics for assessing rendering quality.

As illustrated in Table 3, our model significantly surpasses all baseline models in both novel view synthesis and semantic segmentation tasks. Notably, when compared to Semantic-Ray, which utilizes a pretrained NeuRay [21] model for initialization, our approach, which incorporates scaled pretraining and a hardness-aware strategy, demonstrates markedly superior generalization capabilities. In Figure 6, comparing with the strongest baseline Semantic-Ray, our NPS has better rendering quality and clearer semantic segmentation.

#### 4.2.2 Multi-View Depth Estimation

**Baselines** Our model is benchmarked against two generalizable NeRF models: VolRecon [28], which is trained for rendering and depth, and MVSNeRF [2], which is enhanced with an additional depth head akin to the semantic head. We also compare our model with the multi-view stereo model MVS2D [44]. Given that MVS2D relies on ground-truth target views for depth map generation and does not generate novel views, we use renderings from our NPS model as the target view input for a fair comparison.

Predicting both the novel view and its corresponding

| Method | PSNR↑ | SSIM↑ | LIPIPS↓ | mIoU↑ | Total Acc↑ | Avg Acc↑ |
|---|---|---|---|---|---|---|
| Semantic-NeRF | 25.07 | 0.797 | 0.196 | 91.24 | 97.54 | 93.89 |
| MVSNeRF+Semantic Head ft | 23.84 | 0.733 | 0.267 | 55.26 | 76.25 | 69.70 |
| NeuRay+Semantic Head ft | 27.22 | 0.840 | 0.138 | 77.48 | 91.56 | 81.04 |
| Semantic-Ray ft | 29.27 | 0.865 | 0.127 | 91.08 | 98.20 | 93.97 |
| Ours ft | **29.53** | **0.873** | **0.119** | **93.12** | **98.46** | **95.26** |

Table 3. Quantitative comparison on scene rendering and multi-view semantic segmentation on ScanNet.

| Method | PSNR↑ | SSIM↑ | LIPIPS↓ | AbsRel↓ | RMSE↓ |
|---|---|---|---|---|---|
| MVS2D | – | – | – | **0.112** | 0.257 |
| MVSNeRF+Depth ft | 22.01 | 0.701 | 0.352 | 0.196 | 0.448 |
| VolRecon ft | 15.31 | 0.572 | 0.593 | 0.145 | 0.319 |
| Ours | **23.26** | **0.754** | **0.312** | 0.119 | **0.249** |

Table 4. Quantitative comparison on scene rendering and multi-view depth estimation on ScanNet.

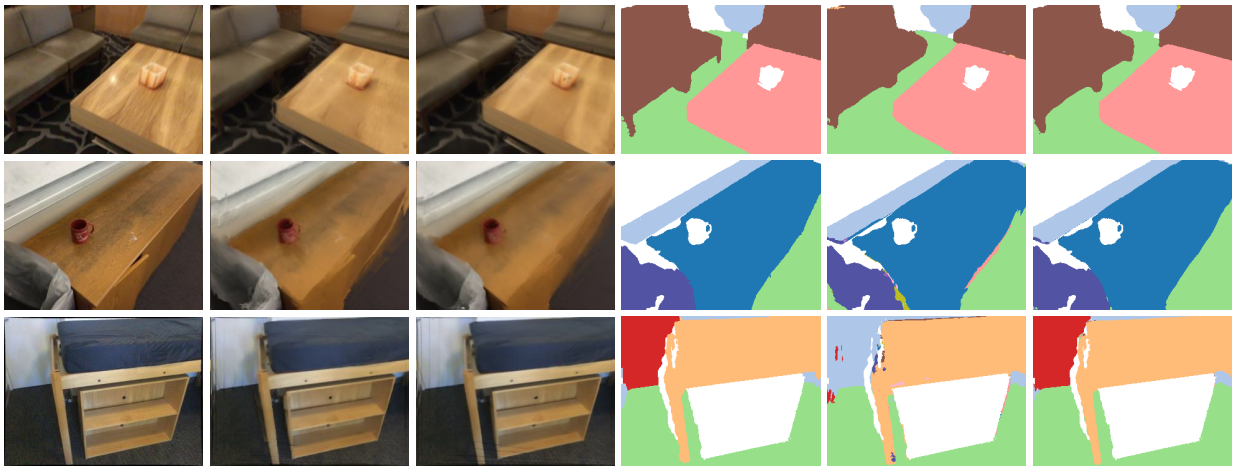



Figure 6. Qualitative results for the multi-view semantic segmentation. The region without ground truth semantic labels is colored in white.

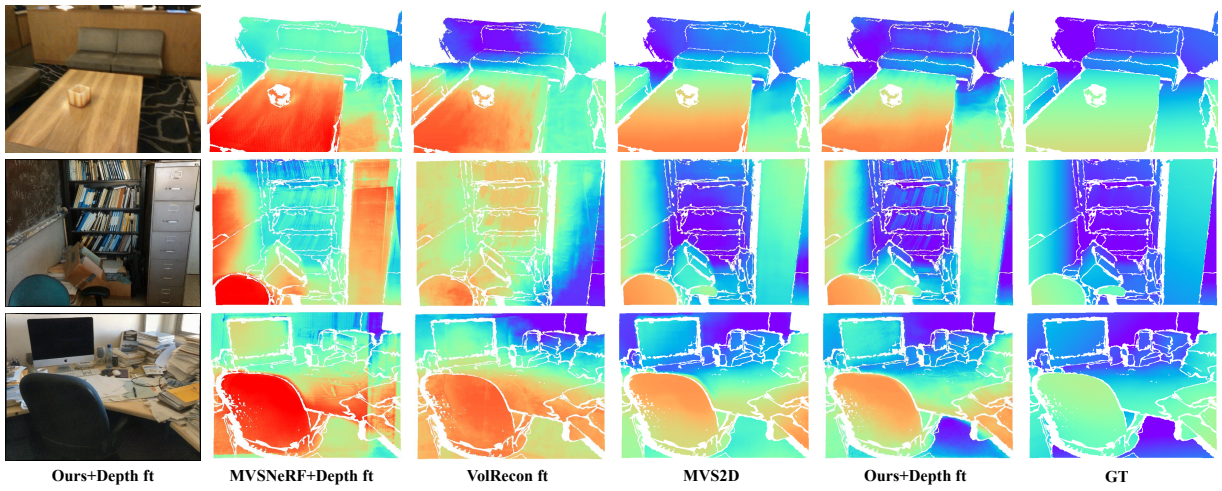



Figure 7. Qualitative results for the multi-view depth estimation. Each row corresponds to one test example. The region without ground truth depth labels is colored in white.

depth map from existing source views poses a significant challenge, as it requires inferring both rendering and geometric aspects. As demonstrated in Table 4, our NPS model significantly outperforms other generalizable methods in both rendering and depth estimation, highlighting the success of our pretraining strategies. It also achieves comparable results to MVS2D in depth estimation. Figure 7 further illustrates the effectiveness of our model, showcasing depth estimations that closely match the ground truth and visually appealing renderings of the novel views.

### 4.3. Ablation Studies

We report ablation analysis on our hardness-aware training strategies. Due to the space limit, we defer them to the supplementary. Overall, our studies suggest that though may impede the novel view synthesis performance to some extent, they could contribute to the generalization of our NPS.

## 5. Conclusion

In this work, we draw inspiration from Masked Language Model (MLM) pretraining in BERT, and propose to scale up the generalizable novel view synthesis (NVS) training objective of generalizable NeRF as a pretraining strategy in multi-view scenario. Our approach, NeRF as Pretraining at Scale (NPS), utilizing a large-scale, minimally annotated dataset and a scaled-up model, not only enhances generalizability in both zero-shot and few-shot novel view synthesis, but also reveals emergence capabilities on downstream tasks like depth estimation and semantic segmentation. This breakthrough underscores NPS's versatility and efficiency, reducing the need for task-specific models and setting a new benchmark in the NeRF field.

## References


[1] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 3

[2] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021. 6, 7

[3] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. Abo: Dataset and benchmarks for real-world 3d object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21126–21136, 2022. 3

[4] Wenyan Cong, Hanxue Liang, Peihao Wang, Zhiwen Fan, Tianlong Chen, Mukund Varma, Yi Wang, and Zhangyang Wang. Enhancing nerf akin to enhancing llms: Generalizable nerf transformer with mixture-of-view-experts. *arXiv preprint arXiv:2308.11793*, 2023. 1, 2, 3, 4, 6

[5] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 6

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1, 3

[7] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022. 3, 4, 6

[8] Zhiwen Fan, Peihao Wang, Yifan Jiang, Xinyu Gong, Dejia Xu, and Zhangyang Wang. Nerf-sos: Any-view self-supervised object segmentation on complex scenes. *arXiv preprint arXiv:2209.08776*, 2022. 3

[9] Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. Masked autoencoders as spatiotemporal learners. *Advances in neural information processing systems*, 35:35946–35958, 2022. 3

[10] John Flynn, Michael Broxton, Paul Debevec, Matthew DuVall, Graham Fyffe, Ryan Overbeck, Noah Snavely, and Richard Tucker. Deepview: View synthesis with learned gradient descent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2367–2376, 2019. 4, 6

[11] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision*, 129:3313–3337, 2021. 3

[12] Xiao Fu, Shangzhan Zhang, Tianrun Chen, Yichong Lu, Lanyun Zhu, Xiaowei Zhou, Andreas Geiger, and Yiyi Liao. Panoptic nerf: 3d-to-2d label transfer for panoptic urban scene segmentation. In *2022 International Conference on 3D Vision (3DV)*, pages 1–11. IEEE, 2022. 3

[13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2, 3

[14] Wonbong Jang and Lourdes Agapito. Codenerf: Disentangled neural radiance fields for object categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12949–12958, 2021. 3

[15] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 406–413, 2014. 3, 4

[16] Jonáš Kulhánek, Erik Derner, Torsten Sattler, and Robert Babuška. Viewformer: Nerf-free neural rendering from few images using transformers. In *European Conference on Computer Vision*, pages 198–216. Springer, 2022. 4

[17] Hanxue Liang, Hehe Fan, Zhiwen Fan, Yi Wang, Tianlong Chen, Yu Cheng, and Zhangyang Wang. Point cloud domain adaptation via masked local 3d structure prediction. In *European Conference on Computer Vision*, pages 156–172. Springer, 2022. 2, 3

[18] Fangfu Liu, Chubin Zhang, Yu Zheng, and Yueqi Duan. Semantic ray: Learning a generalizable semantic field with cross-reprojection attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17386–17396, 2023. 3, 6

[19] Liu Liu, Wenqiang Xu, Haoyuan Fu, Sucheng Qian, Qiaojun Yu, Yang Han, and Cewu Lu. Akb-48: A real-world articulated object knowledge base. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14809–14818, 2022. 3

[20] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 3

[21] Yuan Liu, Sida Peng, Lingjie Liu, Qianqian Wang, Peng Wang, Christian Theobalt, Xiaowei Zhou, and Wenping Wang. Neural rays for occlusion-aware image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7824–7833, 2022. 6, 7

[22] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019. 4, 6

[23] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1, 6

[24] Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *European conference on computer vision*, pages 604–621. Springer, 2022. 2, 3

[25] Navneet Paul. Transnerf-improving neural radiance fields using transfer learning for efficient scene reconstruction. Master's thesis, University of Twente, 2021. 4

[26] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *European Conference on Computer Vision*, pages 523–540. Springer, 2020. 3

[27] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10901–10911, 2021. 3

[28] Yufan Ren, Tong Zhang, Marc Pollefeys, Sabine Süsstrunk, and Fangjinhua Wang. Volrecon: Volume rendering of signed ray distance functions for generalizable multi-view reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16685–16695, 2023. 2, 3, 7

[29] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 3

[30] Vincent Sitzmann, Semon Rezchikov, William T. Freeman, Joshua B. Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation rendering. In *Proc. NeurIPS*, 2021. 3

[31] Mohammed Suhail, Carlos Esteves, Leonid Sigal, and Ameesh Makadia. Generalizable patch-based neural rendering. In *European Conference on Computer Vision*. Springer, 2022. 1, 6

[32] Mohammed Suhail, Carlos Esteves, Leonid Sigal, and Ameesh Makadia. Generalizable patch-based neural rendering. In *European Conference on Computer Vision*, pages 156–174. Springer, 2022. 3, 4

[33] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022. 3

[34] Alex Trevithick and Bo Yang. Grf: Learning a general radiance field for 3d representation and rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15182–15192, 2021. 3

[35] Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*, 2019. 3

[36] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1588–1597, 2019. 3

[37] Suhani Vora, Noha Radwan, Klaus Greff, Henning Meyer, Kyle Genova, Mehdi SM Sajjadi, Etienne Pot, Andrea Tagliasacchi, and Daniel Duckworth. Nesf: Neural semantic fields for generalizable semantic segmentation of 3d scenes. *arXiv preprint arXiv:2111.13260*, 2021. 3

[38] Peihao Wang, Xuxi Chen, Tianlong Chen, Subhashini Venugopalan, Zhangyang Wang, et al. Is attention all nerf needs? *arXiv preprint arXiv:2207.13298*, 2022. 1, 2, 3, 4, 6

[39] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2021. 1, 3, 6

[40] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14668–14678, 2022. 3

[41] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 803–814, 2023. 2, 6

[42] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015. 3

[43] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019. 3

[44] Zhenpei Yang, Zhile Ren, Qi Shan, and Qixing Huang. Mvs2d: Efficient multi-view stereo via attention-driven 2d convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8574–8584, 2022. 7

[45] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pages 767–783, 2018. 1

[46] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1790–1799, 2020. 3

[47] Jianglong Ye, Naiyan Wang, and Xiaolong Wang. Featurenerf: Learning generalizable nerfs by distilling foundation models. *arXiv preprint arXiv:2303.12786*, 2023. 3

[48] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images, 2020. 3

[49] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 1, 6

[50] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew Davison. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 3, 7

[51] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018. 4, 6